We thank reviewers for their thoughtful comments and provide responses to, in our opinion, the most prominent ones. **[R1, R2] Zero inflation and model misspecification.** We now tested ZIPBN on different percentages of zeros (0%, 25%, 50%, 75%) using otherwise the same setting as in Tab. 2. ZIPBN outperformed MRS, e.g., MCCs were 0.74 vs 0.49, 0.80 vs 0.67, 0.83 vs 0.66, and 0.69 vs 0.34. We also tested ZIPBN on misspecified models, namely, zero-inflated negative binomial BN with dispersion parameters generated uniformly in $(1, 5)$ and 50% zeros. ZIPBN still outperformed MRS with a smaller gap (MCC 0.54 vs 0.46). We anticipate MRS to outperform ZIPBN as %zeros→0. **[All] Real data validation**. It is difficult to directly quantify and compare the estimation accuracy for real large networks due to unknown true structure, which motivated our first real data analysis with transcription factors. For comparison, we now applied the closest competitor MRS to the same dataset, which correctly identified 198 pairs (we identified 304) out of 479. For the pathway analysis, we now computed AIC for ZIPBN and MRS. Because MRS only provides an estimate of the graph, we fit a zero-inflated Poisson regression to each node given its parents estimated from MRS. ZIPBN outperformed MRS with AIC 84840.13 vs 85142.73 (the two graphs have similar sparsity). **[R3] Data filtering.** We agree that filtering out genes with >70% zeros makes the task easier. We did so mainly because genes with >70% zeros also tend to have very low non-zero counts and therefore have extremely small variability across samples (e.g. in our study, the median variance of genes with >70% zeros is 10^{-4}). For that reason, removing those uniformly low expressed genes seems to be a common practice in genomic studies. Note that with the filtering, the remaining genes still have about 50% zeros and low variability (median variance is 1.62), for which our method worked better than competing methods (see the previous paragraph). The purpose of this study is not to prove that ZIPBN works well in extremely sparse data but rather to provide real-world evidence in addition to simulations that when the zero-inflation is moderately large (<70%), ZIPBN is capable of and superior in identifying the correct DAG. **[R3] Causal sufficiency.** Causal sufficiency should have been stated explicitly. We are also aware that papers addressing latent confounders (e.g., Spirtes and Richardson 2002, Salehkaleybar et al. 2020) have been extensively studied. We had focused on the literature most closely related to ours due to the limited space. In terms of methodological development, we focused on developing the first BN model for sparse count data, establishing identifiability theory, and designing effective structural learning algorithm. Introducing latent factors that account for latent confounders would necessarily complicate the model, theory, and computation, and may obscure the contributions of different components (zero-inflation, sparsity prior, tempered MCMC algorithm) of our method to the good empirical performance. Focusing on cases without confounders allowed us to compare with competing methods for count data that do not account for confounders. That being said, it will be very interesting to extend our current work in this direction. **[R3] Existing greedy algorithms.** We agree that Chickering's paper is indeed very relevant to this paper and should have been mentioned in the paper. There are, however, several fundamental differences between the convergence of our algorithm and that of Chickering's, which led us to the claim "unlike heuristic/greedy search algorithms, MCMC is theoretically guaranteed to converge to its stationary distribution". (1) Our algorithm requires a sufficient number of MCMC iterations whereas Chickering's requires a large sample size. In practice, a large enough sample size (relatively to the super-exponential size of DAG space) is often infeasible and expensive to obtain, while it is comparatively much easier and cheaper to increase the size of MCMC and its (lack of) practical convergence can be monitored by various diagnostics (e.g. Gelman-Rubin's potential scale reduction factor). (2) Chickering's method requires faithfulness whereas ours doesn't. Faithfulness can be violated with a limited sample size (see e.g., Ulher et al., 2013 *Ann. Stat.*) and/or in an equilibrium-maintaining system such as a biological system. Intuitively, while both Gaussian and multinomial can have "accidental" cancellation of positive and negative effects (e.g., in a feedforward loop of exercise, body temperature, and sweating) and hence become unfaithful, the proposed ZIPBN doesn't allow such cancellation because of its count nature (also see e.g., Park and Park 2019, *AISTATS*). (3) We focus on identifying individual BN whereas Chickering's focuses on identifying Markov equivalence class. (4) Greedy search algorithm converges to a point estimate whereas MCMC-based method converges to the posterior distribution which allows for finite-sample statistical inference (e.g., edge inclusion probability, FDR control) of the estimated BN. **[R3] Non-sparse prior.** A common assumption for BN structure learning is sparsity, which is induced via sparse graph priors in ZIPBN and sparse skeletons in ODS/MRS. For all methods, the assumed sparsity level can significantly affect performance. ZIPBN and many other Bayesian methods do not use a uniform graph prior (a special case of Erdös-Rényi with probability 0.5) because it favors dense graphs: (i) Erdös-Rényi is a well-known dense random graph and (ii) there are far more dense graphs $|\boldsymbol{E}| = O(p^2)$ than sparse graphs $|\boldsymbol{E}| = O(p)$ (e.g., a graph with half the size of a complete graph is still very dense). So as expected, when a dense uniform prior was adopted, FDR of ZIPBN went from 0.18 to 0.59 (n=500,p=50). However, same goes for ODS/MRS when skeleton learning cutoff is chosen to favor dense graphs. **[All] Computation speed.** The worst-case per iteration cost is $O(np^2)$ (mainly the likelihood evaluation), which is reduced to $O(\max(n, p)p)$ for sparse networks (i.e., $|\boldsymbol{E}| = O(p)$). The CPU time for real pathway analysis was 1.7hrs. **[All] Clarifications.** p_s was chosen to be 10% and the algorithm appears not sensitive in that choice. We gently point out that Theorem 1 does hold for *all* networks because we proved that *any* two Markov equivalent networks with different structures (i.e., $\boldsymbol{E'} \neq \boldsymbol{E}$, line 3 of the proof) have to have different distributions (note that two non-Markov-equivalent BNs are trivially not distribution equivalent, e.g., Spirtes et al. 2000). FDR was controlled at 1% (line 286). We will improve the description of priors, MCMC, intuition of non-identifiable BNs, notations, and broader impacts.